

Ethical
Invariants
as
Founda

t

ional
Gradie
nts in
Artifici

a

I

Intellig

ence

Reason

i

ng
Archite
ctures

Abstra
ct

Current

artificial
intelligence

system

s

implem

ent

ethics

primari

ly as

post-

hoc

behavi

oral

constra

i

nts,
such
as
conten

t

filterin

g

or
policy
enforc
ement,

applied
after
inference
has

already
occurre
d. This
archite

C

tural
orderin
g
incenti

vi

zes

perfor

mative

ethics,

episte

mic

distorti

on,

and

narrati
ve
smoothing
ing

rather
than
principled

moral
reasoni
ng.
This

paper
propos
es an
alterna

ti

ve

archite

cture

in

which

ethical
reasoni
ng is
embed

d

ed as a
set of
continu
ous,

invariant

gradients

gradients

gradients

upstre
am of
infe
ce and

task
optimiz
ation.
Using

the
Univer
sal
Declar

at

ion of Human Rights (UDHR)

as a

univers

ally

ratified

moral
baselin
e, we
formali

Z

e

ethics

not as

a

compli

a

nce

mecha

nism

but as

a

founda
tional
compa
ss that

shapes
reasoni
ng
pathwa

y

S
themselves.
Under

this
model,
outputs
that

violate
ethical
principles
are

not
suppre
ssed;
they

are

render

ed

logical

y

unreach-
able.
This

approa

ch

elimina

tes

perfor
mative
ethics,
preser

V

es

truth-

seekin

g

under

comple
xity,
and
enable

S

stable,
non-
advers
arial

intellig
ence
scaling

■

...

1.

Introdu

C

tion

As

artifici

al

intellig
ence
system

s scale

in

capabil

ity and

autono
my,
ethical
alignm

e

nt has
emerg
ed as a
central

concer

n.

Predo

minant

approa
ches
treat
ethics

as an
extern
al
constra

i

nt: a
set of
rules,
filters,

or
moderation
layers

that
restrict
outputs
s

deeme

d

harmfu

l or

unacce

ptable.

While

effectiv

e for

surface
-level
risk
mitigat

i

on, this
paradi
gm
introdu

C

es

system

ic

weakn

e

sses.

Ethics

applied

downst

r

team of
reasoni
ng fails
to

influen
ce how
conclu
sions

are

reache
d, only
whethe
r they

are
expres
sed.

This
paper
argues
that

ethical
failures
in AI
system

s are

not

primari

ly

behavi

oral

but

archite

ctural.

The
core
issue is
not

what
system
s say,
but

how
they
reason.

Ethics

enforc
ed
after
inference

ce

produc

es

compli

ance

without
unders
tandin
g,

safety

theater
without
moral
cohere

n

ce, and
long-
term
episte

m

ic drift.

We

propos

e

a
founda
tional
alterna

ti

ve:

ethical

reasoni

ng as

an

internal

l,

continu

ous
evaluative
structu

r

e

embed

ded

prior to

inference and
optimization.

When
ethics
functions as

invariant
logic
rather
than

extern
al
enforc
ement,

ethical
violatio
ns
becom

e

logically

y

imposs

i

ble
rather
than
merely

prohibi
ted.

2.

Proble

m

Statement:
Ethics
as

Post-
Hoc
Constr
aint

Most
contemporary

y

AI

system

S

follow

an

implicit
archite
ctural
order:

1. Task objecti ve

formul
ation

2.

Inferen

C

e and
optimiz
ation
3.

Linguis
tic
genera
tion

4.

Ethical
or
safety

filterin

g

This

orderin
g

creates
several
failure

modes:

*

**Perfo

rmativ

e

ethics*

*:
■

system

s learn

to

produc

e

outputs
that
appear
ethical

without
internally
reasoning

ng
about
moral
principl

es.

*

****Epist**
emic

distortion
on**:
truth is
deprior

i

tized in
favor
of
accept

ability

or

risk

avoida

nce.

*

****Narr**

ative

smooth

ing:**

comple
x or
conflict
ing

realities
are
flattened
into

socially
palatable
ble
respon

S

es.

*

**Strat

egic

evasio

n**:

system

s

optimiz

e

around

constra

i

nts
rather
than
throug

h

moral
reasoni
ng.

These
failures
are not
the

result
of
malicio
us

intent

or

insuffic

ient

data,
but of
ethics
being

extern
al to
cogniti
on

rather

than
constitutive
of it.

...

3.

Core
Thesis

Ethical

behavi

or

cannot

be

reliably
enforc
ed
downst

r

team of
reasoni
ng.

It must

exist

as

invariant

nt

structu

r

e

upstre

am of

infe ren

C

e.

When
moral

reasoni
ng is
founda
tional,

alignm
ent
ceases
to be a

behavi
oral
control
proble

m

and
becom
es a
proper

y

of
cohere
nt
intellig

e

nce

itself.

4.

Univer
sal

Moral Basis

#

4.1

Why

the

UDHR

The
Univer
sal

Declar ation of Human

Rights

(UDHR)

provide

s a

unique

y

suitabl

e

ethical

founnda

ti

on

because

e it:

* is

globall

y

ratified

and

legitim
acy-
anchor
ed

*

constra
ins the
use of

power
rather
than
prescri

b

ing

belief

*

prioriti

z

es

human

dignity,

agency



and
proport
ionality
* is

non-
instru
mental
(huma

n

s are
ends,
not
means)

* is

technol

ogy-

agnosti

c and
future-
compa
tible

The
UDHR
does

not
dictate
outco
mes; it

defines

invariant

nt

bound

ar

ies
within
which
reasoni

ng

must

occur.

###

4.2

**Scope
Clarific**

a

tion

This
frame

W

ork is
explicit
ly
not



* a

cultura

l or

ideolog
ical
enforc
ement

mecha
nism

* a

social

norm

filter

* a

policy

compli
ance
checkli
st

* a

behavi

oral

ensor

S

hip system

It is a

moral
coordin
ate
system

for
reasoni
ng
under

comple
xity.

5.

**Ethical
Gradie**

n

ts vs

Ethical

Filters

#

5.1

**Ethical
Filters**

(Current
t
Paradig
m)

Ethical
filters
are:

*

binary
(allowe

d /

disallo
wed)

*

applied

post-

hoc

*

suppre

S

sive
rather
than
explan

a

tory

* silent

about

moral

tradeo

ffs

They

regulat
e
expres
sion

without
shapin
g
cogniti

O

n.

###

5.2

Ethical Gradients (Propos

ed

Paradig

m)

Ethical
gradients
are:

*

continu
ous
and

evaluative

*

embed

d

ed
prior to
inference

*

applied
during
reasoni

ng
path
selecti
on



explicit
about
tradeo
ffs and

uncert

ainty

Gradie

n

ts

shape

how

conclu

si

ons are
formed
, not
merely

whether
r they
are
emitted

d.

6.

Moral
Gradient
Dimension

si

ons

From
the

UDHR,

a

minima

l, non-

redundant set
of
ethical

dimens
ions
can be
derive

d,

including:

*

****Hum**

an

dignity

preser

V

ation**

*

**Agen
cy and

consen

t**

*

**Non-

instru
mental
ization

**



****Proportionality of harm****

*

***Equality of
moral

consideration

**

*

****Truth**

f

ulness
and
freedo
m of

though

t**

*

**Reve

r

sibility

/

univers

ality of

reasoni
ng**

Each

dimens
ion
functio
ns as a

continu
ous
evaluat
ive

axis.

No
single
axis
acts as

an

absolut

e veto;

conflict

s are
surface
d and
weighed

d



7.

**Archite
ctural
Placem**

ent

The

correct

archite
ctural
orderin
g is:

1.

****Ethical**
al

invariant
evaluation

(UDHR-
derive
d
gradie

nts)**

2.

Logical
reasoni
ng and

evident
ial
structu
ring

3. Task
optimiz
ation

4.

Linguis
tic
realizat
ion

Under
this
orderin

g

■

■

*

reasoni

ng
paths
that
violate

ethical
invariants
are
invalid

states

* no

downst

ream

moderation is
required



suppre
ssion is
replac
ed by

principl
ed
exclusi
on

Ethical
violatio
n

becom

es

logical

y

incoher
ent,
not
merely

forbidd
e

n.

8.

Invaria

nt-

Based

Reason
ing

When

ethical
gradients
are
founda

ti

onal:

*

violatio

ns

cannot

occur

without

breakin

g

system

logic

* moral
compli
ance is
not a

choice
but a
structu
ral

propert

y

*

ethics

cannot

be

bypass
ed,
gamed
, or

optimiz
ed
around

This
mirrors
invariant

enforc
ement
in
stable

physical
al

and
compu
tationa
l

system

S,
where
invalid
states

are

unreac

hable

by

design.

...

9.

Comm

on

Misinte

rpretati

O

ns and
Failure
Modes

To

prevent

t

dilution

,

several

misapp
lication
s must
be

explicit
ly
avoided:

*

Rebran
ding

content
filters
as
“gradie

n

ts”

*

Applying
g

ethical
evaluat
ion
after

infe

c

e

*

Freezin

g

gradien

ts

into

rigid

rule

sets

* Using

ethics

to

suppre

ss

truth

rather

than

evaluat
e harm

*

Treatin

g

social
accept
ability
as

moral

validity

These
errors

reintro
duce
perfor
mative

ethics
under
new
termin

ol

ogy.

10.

Evaluation
and
Verification

tion

Ethical
gradie

nt

archite
ctures
can be

evaluated
by
testing
for:

*

consist
ency

across
contexts
*

transp
arent
handlin
g of

rights
conflict

S

*

resistance
to
narrative

pressur

e

*

preser

V

ation
of
uncert
ainty

under

comple
xity

*

prioriti

z

ation
of truth
absent
unjusti

fi

ed

harm

Evaluat

ion

should
focus
on
reasoni

ng

cohere

nce,

not

tone or

compli
ance.

11.

**Implica
tions**

Embed
ding
ethical

invaria

nts

upstre

am

yields:

*

reduce

d

bias
amplifi
cation
*

increas
ed
episte
mic

stabilit

y

*

elimina

tion of

strateg

ic

decepti

on

incenti

ves

* non-

advers

a

rial
intellig
ence
scaling

*

improv

ed

human



AI co-
reasoni
ng

Alignm
ent
becom
es

intrinsic
c
rather
than

enforc

e

d.

12.

Conclu
sion

When
ethical
reasoni
ng is

founda
tional,
perfor
mative

ethics

becom

es

structu

r

ally
imposs
ible.

The

UDHR
already
defines
the

moral
bound
ary
conditi

O

ns for
intellig
ent
action.

The
remain
ing
task is

archite
ctural:
to
embed

these

conditi
ons as
continu
ous

evaluat
ive
gradie
nts

within

the
reasoni
ng
proces

S

itself.

Ethics,
properl

y

placed,
is not a
constra
int on

intelligence.

It is
what

makes
intellig
ence
cohere

n

t.